

Exploring the Current Practices, Costs and Benefits of FAIR Implementation in Pharmaceutical Research and Development: A Qualitative Interview Study

Ebtisam Alharbi^{1,2}, Rigina Skeva¹, Nick Juty¹, Caroline Jay¹ & Carole Goble^{1†}

¹School of Computer Science, University of Manchester, Manchester, Manchester M13 9PL, UK

²College of Computer and Information Systems, Umm Al-Qura University, Mecca, Makkah 21421, Saudi Arabia

chinaXiv:202211.000202v1

Keywords: FAIR; FAIRification; Retrospective FAIRification; Pharmaceutical R&D; Cost-benefit; Decision-making process

Citation: Alharbi, E., et al.: Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical research and development: A qualitative interview study. *Data Intelligence* 3(4), 507-527 (2021). doi: 10.1162/dint_a_00109

Received: April 14, 2021; Revised: September 21, 2021; Accepted: September 23, 2021

ABSTRACT

The findable, accessible, interoperable, reusable (FAIR) principles for scientific data management and stewardship aim to facilitate data reuse at scale by both humans and machines. Research and development (R&D) in the pharmaceutical industry is becoming increasingly data driven, but managing its data assets according to FAIR principles remains costly and challenging. To date, little scientific evidence exists about how FAIR is currently implemented in practice, what its associated costs and benefits are, and how decisions are made about the retrospective FAIRification of data sets in pharmaceutical R&D. This paper reports the results of semi-structured interviews with 14 pharmaceutical professionals who participate in various stages of drug R&D in seven pharmaceutical businesses. Inductive thematic analysis identified three primary themes of the benefits and costs of FAIRification, and the elements that influence the decision-making process for FAIRifying legacy data sets. Participants collectively acknowledged the potential contribution of FAIRification to data reusability in diverse research domains and the subsequent potential for cost-savings. Implementation costs, however, were still considered a barrier by participants, with the need for considerable expenditure in terms of resources, and cultural change. How decisions were made about FAIRification was influenced by legal and ethical considerations, management commitment, and data prioritisation. The findings have significant implications for those in the pharmaceutical R&D industry who are engaged in driving FAIR implementation, and for external parties who seek to better understand existing practices and challenges.

[†] Corresponding author: Carole Goble (Email: carole.goble@manchester.ac.uk; ORCID: 0000-0003-1219-2137).

1. INTRODUCTION

The FAIR principles articulate the importance of making scientific research data findable, accessible, interoperable, and reusable (FAIR) [1]. These principles have been initiated in a few academic institutions and have thus gained popularity and endorsement [2], leading to their wide acceptance by policymakers [3], funding councils [4], publishers [5], and research communities [6]. The ultimate goal of these aspirational principles is to enhance the data infrastructure by enabling data reuse at scale by both humans and machines [7]. It has been estimated that not having FAIR research data costs the European economy at least €10.2 billion per year [8]. As such, governments [9], international [10], and local [11] organisations are using the FAIR principles to drive data management strategy in both the public and private sectors.

Seeing the potential of implementing FAIR principles, the pharmaceutical industry has responded quickly [12] and is to tackle the data challenges faced by these large, complex global enterprises [13]. Implementing these principles as effective data management strategies could amplify the value of data assets through higher data reusability [14]. As pharmaceutical research and development (R&D) is increasingly becoming a data-driven process, significant effort must be devoted to managing its data assets efficiently and effectively. Over the past two decades, the cost of drug R&D has risen ten-fold, whereas the number of approved new drugs has steadily declined [15, 16]. For many years, the Innovative Medicines Initiative (IMI)^① has sponsored data management projects that have dealt with developing data centres. These projects have shown that proper data asset annotation and management is a complex, resource-intensive process that must be improved [17, 18, 19, 20].

Progress has been made towards the adoption of these principles, led by the Pistoia Alliance FAIR toolkit^②, a pharmaceutical company collaboration in the pre-competitive space that aims to facilitate FAIR implementation [21], and the FAIR cookbook from the IMI FAIRplus project^③. The latter is an activity of the FAIRplus project^④, an ongoing EU project to develop tools and guidelines for making data FAIR in collaboration with the European Federation of Pharmaceutical Industries and Associations (EFPIA) [22]. These initiatives, which play a significant role in transforming data management and stewardship, make a concerted effort to drive FAIR implementation in pharmaceutical R&D.

Previous research has shown that adopting FAIR guidelines for data management has the potential to increase the efficacy of drug R&D [12, 13]. More specifically, studies reported that the availability of FAIR data for its original purpose and beyond (primary and secondary use) can accelerate innovation and reduce the time needed to bring a drug to market [12]. Furthermore, this improvement in the discovery and development of innovative medicines has been driven by the exploitation of advanced analytical technologies, such as artificial intelligence (AI) and machine learning [23, 24, 25, 26].

^① <https://www.imi.europa.eu>

^② <https://fairtoolkit.pistoiaalliance.org>

^③ <https://github.com/FAIRplus/the-fair-cookbook>

^④ <https://fairplus-project.eu>

Despite the potential that the FAIR principles offer pharmaceutical R&D, their implementation poses significant challenges. Existing research has briefly highlighted the obstacles that might impact the effective implementation of FAIR at an enterprise level [12, 27]. A lack of financial investment, technical infrastructure, training, and cultural change were the most commonly identified barriers. However, it is balancing the requirements of diverse stakeholders involved in the R&D, enterprise, IT, and business domains that presents the most significant challenge [3, 12, 13]. A recent study indicated that most pharmaceutical companies were at an early stage of internal FAIRification, focused on the process of aligning data sets with FAIR principles [28], which is often driven by use cases due to these challenges [13].

Retrospective FAIRification—making legacy data sets align with FAIR principles—offers significant potential, but this also remains limited [29]. A recent report found that the reason FAIR is hard to achieve at scale in the pharmaceutical industry is due to the challenge of dealing with existing legacy data [30, 31]. A key challenge in retrospective FAIRification is the cost, which includes the upfront cost of revising legacy data to comply with data standards, the previous investment in legacy systems, and the cost of data loss during transformation [27].

Although some previous research has discussed FAIR implementation in the context of pharmaceutical R&D [12, 13], there is little study of its actual implementation in a company setting. In this paper, we examine current approaches to FAIR implementation in pharmaceutical R&D using a qualitative approach that explores the experiences of pharmaceutical professionals. This paper will be of interest to those in the pharmaceutical industry who are engaged in FAIR implementation and to external parties who seek to better understand existing practices and challenges. Ultimately, the results will be used to develop a decision-making framework to aid decision makers in pharmaceutical R&D to determine whether FAIRifying a legacy dataset is worth the cost of the investment, and to help them prioritise their data sets accordingly.

2. METHOD

In this study, we conducted semi-structured interviews to gain deep insights into the current implementation of FAIR data principles in pharmaceutical R&D. Semi-structured interviews effectively capture the complexities of a phenomenon and enable further in-depth exploration of experts' thoughts in an open-ended manner [32, 33, 34]. The interviews aimed to comprehensively explore the thoughts of the experts involved in FAIR implementation, and covered the associated costs and expected benefits, and how decisions were made about the retrospective FAIRification of data in pharmaceutical R&D.

2.1 Participants

We recruited 14 participants (4 females and 10 males) working in pharmaceutical companies involved in the implementation of the FAIR data principles. The sampling used both purposive and snowball techniques [35]. The inclusion criterion was at least two years' experience handling life science data and working with FAIR guiding principles. Participants were recruited from the European Federation of Pharmaceutical Industries and Associations (EFPIA) members participating in the FAIRplus project. Eligible

participants were sent invitations to take part in online interviews using the email address that appeared on their personal pages or those provided by the FAIRplus project team. Table 1 summarises the participant profiles, including their role in their companies and their area of expertise.

Table 1. Summary of participant information.

ID	Role	Area of expertise	Experience years
P1	Data manager	Pre-clinical research	10–15
P2	Assistant head	Data and knowledge management in R&D	10–15
P3	Data director	Data management, data science, and AI in R&D	20–25
P4	Data curator	Bioinformatics, identifiers, and data hosting	5–10
P5	Principal IT business manager	Clinical pharmacology and Safety Sciences	20–25
P6	Technical assoc. director	Data ontology and mapping domain	10–15
P7	Alliance manager	Drug development and biomarker research	25–30
P8	Data manager	Life science informatics and drug discovery	15–20
P9	Manager of discovery programmes	Data curation across biopharma and functional genomics	1–5
P10	Member of data strategy team	Ontologies, standardisation processes, curation, data strategy and FAIR definition	5–10
P11	Director	Bioinformatician in neuroscience	1–5
P12	Principal analyst	Data curation of clinical and preclinical studies	10–15
P13	Principal scientist	Data management plans and project sustainability	5–10
P14	Senior director	Drug discovery, development, manufacture and commercialization	15–20

2.2 Procedures

All participants provided informed consent prior to taking part in the study. Each was interviewed once online (via Zoom) by a single researcher (the first author). The interview started with a brief introduction to the study. All the interviews were audio-recorded, transcribed and anonymised. Each of the sessions lasted between 30 and 60 minutes. The interview questions were used as prompts for the discussions, which varied in terms of detail depending on the role of an interviewee. The interview guide covered the following questions:

1. What are the current FAIR data practices in your company?
2. What are the motivations for your company's FAIRification programme?
3. What kind of FAIRification are you targeting—prospective or retrospective FAIRification of data sets?
4. What are the activities involved in FAIRification?
5. What is needed in terms of resources to implement FAIRification activities, and why?
6. Who are the stakeholders involved in FAIRification?
7. What are the costs associated with FAIRification?
8. What are the benefits of FAIRifying a data set to your company?
9. Which parts of the drug discovery value chain are more important for FAIRification than others?

10. What are the reasons you decided against FAIRifying a legacy data set?
11. What is your process of selecting a data set for FAIRification?
12. How is the decision to FAIRify made?
13. Is there any evidence that FAIRifying a data set returns value? Please give examples.

2.3 Analysis

The interview transcripts were uploaded to the qualitative data software NVivo 12[®], and were thematically analysed [36]. The themes were identified using an inductive method (open coding) using the following steps: (1) repeated reading of the transcripts by the first author for familiarisation with the content; (2) initial ideas converted into relevant concepts–codes; (3) preliminary codes identified to contextualise the data; and (4) themes reviewed iteratively until each code was effectively represented by the extracts attached to it. Finally, an independent coder (second author) who was not involved in the study design or theme generation was given the codebook and transcripts to test the reliability of the coding. The inter-coder reliability analysis of the transcribed interviews yielded a percentage agreement of 79.1% and Cohen's kappa (κ) of 0.66, which indicates substantial agreement.

2.4 Ethical Approval

The study was granted ethical approval by the Research Ethics Committee of the University of Manchester (Ref.: 2019-7982-12464).

3. RESULTS

3.1 Themes

The thematic analysis identified three primary themes: FAIRification benefits, FAIRification costs and the FAIRification decision-making process. Each theme, along with relevant sub-themes (Figure 1) is described in further detail below.

[®] <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software>

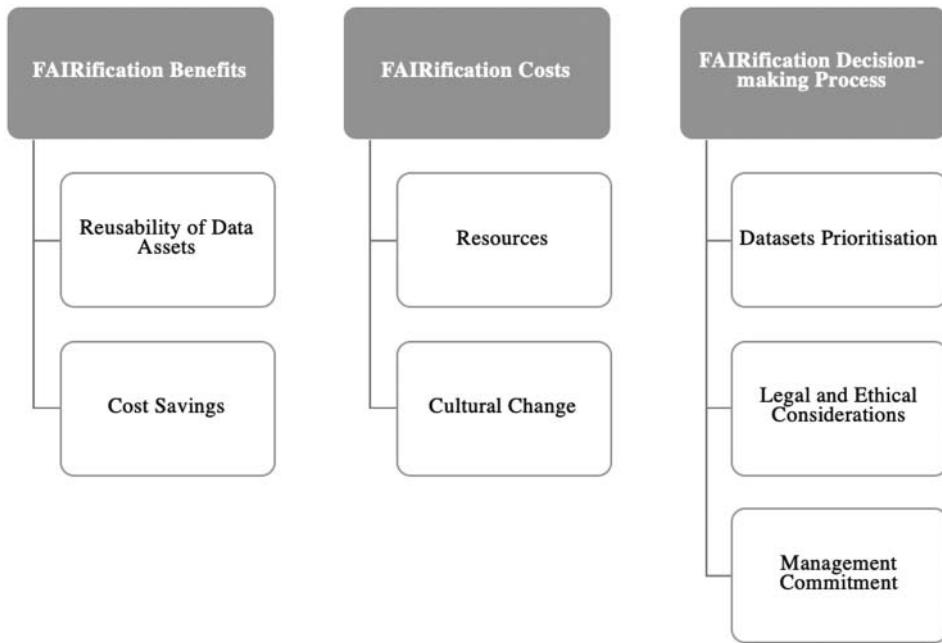


Figure 1. The thematic analysis themes and sub-themes.

The perspective of the pharmaceutical professionals was that FAIR data stewardship should be considered a corporate data management strategy, important for improving efficiency. FAIR implementation was viewed as particularly important for pharmaceutical R&D due to the complex and disconnected data landscape. The participants emphasised that the FAIR data principles would address several issues inherent in their company settings, such as siloed, project-based data and constant changes in knowledge and expertise. They saw working in a FAIR environment as breaking these siloes, facilitating data sharing practices and ensuring business continuity.

Participants emphasised that implementing FAIR principles is a new practice in their pharmaceutical organisations; community practices are still developing, and growing knowledge about FAIR implementation is still in the process of being assimilated by pharmaceutical R&D units. A few companies have initiated data FAIRification projects on a larger scale, but most are considering data FAIRification in the context of specific use cases.

3.1.1 Theme 1: FAIRification Benefits

This theme describes the benefits expected from implementing FAIR principles in pharmaceutical R&D. The reusability of data assets at scale was identified by the participants as the main benefit. This process was seen as useful in generating value from data assets by enabling companies to utilise the data to create novel scientific insights through facilitative use of advanced analytical approaches, such as AI. The expected financial impact in terms of cost savings and time was also discussed.

(1) Reusability of data assets

The reusability of data was considered the main advantage of implementing FAIR principles in pharmaceutical R&D. The participants stated that they have an enormous amount of legacy data and want to utilise and repurpose those assets to exploit their full value. They explained there are teams specifically concerned with historical data and attempting to convert the data to align with FAIR principles.

“We have loads and loads of legacy data. We would, as much as possible, like to utilise those data as well. That is why there are teams dealing with those legacy data and trying to transform those data such that they fulfil the FAIR requirements.” (P1)

The participants also emphasised that reusing previously generated data has long-term benefits for pharmaceutical R&D, particularly in disease-related areas in which legacy data may offer alternative indications for a drug that companies already have—positive or negative aspects that they may not have realised. As an example, one participant mentioned the response to the COVID-19 pandemic, which would have benefitted from aligning SARS-CoV-2 data with FAIR principles.

“.... it’s driven by a massive societal issue. We are desperately trying to go back and look at what we knew from SARS 10 years ago.” (P8)

The participants stated that the application of FAIR principles could create “future-proofing” and thereby enable rapid innovation. They emphasised that the availability of data in a FAIR format would enable large-scale analysis and the use of innovative, AI analysis methods, such as machine learning techniques.

“The data are in a form that can be cut and diced based on the questions that are being asked rather than the original preformed hypothesis that’s being tested. You open up the doors for machine learning and artificial intelligence.” (P5)

Although the ultimate goal of FAIR implementation is reusability, the participants also noted that improved findability would add a tangible benefit to their businesses, as finding data sets of interest is currently a huge issue in large and complex pharmaceutical organisations. They reported that their data and infrastructure is fragmented across many departments and the simple ability to find what already exists would be extremely beneficial.

“We are still back at the F of FAIR. I think just finding the data would be a big win. For people to find a study, to be able to find all studies across the company with a certain compound or a certain disease would be very useful.” (P6)

An added advantage is that ensuring compliance with FAIR principles presents real value in facilitating data integration. The participants stated that rendering existing data interoperable would improve their ability to integrate large volumes of data and validate results.

“The value of it is the ability to integrate. I want to have more data. I think biologists also recognize that. They want to be able to look across and compare their data with others to see if you get similar results or contrasting results.” (P11)

(2) Cost savings

Aligning data with FAIR principles would have a positive financial impact on pharmaceutical organisations, as it would enable them to maximise value from their data assets. They explained that the availability of relevant data can prevent the duplication of experiments, which in turn, lowers costs and accelerates timelines across the R&D pipeline.

“I can see that it also benefits on the financial side in that if you have a fully FAIR system, then you should be able to avoid redundancies in experiments.” (P2)

Data scientists were identified as primary beneficiaries of implementing FAIR principles, as the availability of FAIR data would save time and money by allowing them to focus on what they considered to be more important, skilled work.

“I think for individual scientists, they spend so much time working on data sets that there is a time and efficiency saving to be achieved if they can easily get a hold of the data sets that they need to do their work. That frees them up to do other more exciting work, writing papers, or going back to the lab.” (P2)

The participants mentioned drug repurposing as an example of reusing existing data in a different way. They stated that repurposing or reusing the same data models or data templates allows for the rapid analysis, transformation or curation of data. This practise helps with identifying the promising drug targets which accordingly minimises costs that accompany the launch of a medicine in the market.

“If you’re able to do target identification quickly or be able to do a drug repurposing. It’s more about saving time and saving costs.” (P12)

3.1.2 Theme 2: FAIRification Costs

This theme centres on the costs associated with implementing FAIR data principles in pharmaceutical R&D departments. Despite the potential future reduction of costs where data have been FAIRified, the FAIRification process itself entails considerable expenditure in terms of resources, both technical and human. Cultural change was also raised as a primary concern in effectively implementing FAIR principles.

(1) Resources

FAIRification was collectively acknowledged by the participants as a resource-intensive task, especially when it was carried out retrospectively. Participants noted that an issue consistently arose: the resources that are available for implementing FAIR principles in their organisations. Resource costs associated with this task included the time, effort, and potentially standing up expenses.

“It is the resource costs of curators and data specialists, data stewards; the resource costs of defining and building metadata models; the implementation costs of things like a reference and master data management.” (P3)

An internally integrated infrastructure was regarded as a requirement of FAIR implementation, due to inconsistency in the existing internal systems. The respondents identified several internal IT applications (e.g., identifier systems, ontology services and storage databases) that support FAIR implementation. However, they also argued that these applications are incompatible with one another and require sophisticated design to achieve effective integration and reconciliation.

“If you don’t have applications in IT infrastructure, so servers, databases and data acquisition pipelines, if that is not all in place, then you have disconnected in the execution of that data capture and analysis interpretation. That creates potential breaks in the FAIR backbone because you don’t have a connected integrated system.” (P5)

For some participants, FAIRification is a distracting task that diverts an individual’s attention from what he/she is supposed to do during a novel scientific investigation. The respondents asserted that the time spent aligning data with FAIR principles might affect an individual’s productivity and thereby significantly influence a company’s day-to-day business of drug discovery. They stated that the FAIRness of the data is not their top priority, but rather a priority secondary to the scientific progression of the project. They declared that individuals and groups in their businesses are assessed on their productivity and research outputs, and not against longer-term objectives such as the extent to which they generate FAIR data.

“People will tell you, ‘I have my setup in place. My objective, I don’t know . . . do another clinical trial, track a new market, this kind of stuff’. They will see this FAIRification as a distraction. This is your typical problem in FAIR data.” (P10)

Additionally, some participants discussed continuity and long-term objectives as essential to the implementation of a FAIRification programme. Within industry, staff churn and organisational change occur frequently. This is a major issue, as personal knowledge plays a significant role in familiarity with the data sets.

“It is a situation where you have a high turnover of staff and a high turnover of expertise. Having the results of work in a FAIR format ensures business continuity. What if individuals leave or for some other reason, or there’s a change in focus which results in a loss of expertise.” (P2)

(2) Cultural change

A pressing issue in pharmaceutical organisations is that the current culture is not conducive to the implementation of FAIR data principles, and that awareness needs to be raised about the importance of FAIR principles to achieve the required cultural change.

“What is needed, as I said, is the cultural change. It is the understanding of their data and it is the understanding that delivering FAIR data does increase the resources to produce those data.” (P1)

Skill sets identified as necessary for data FAIRification were related to eight distinct types of knowledge or abilities, namely, ontologies, metadata, data analysis, data stewardship, domain knowledge, software,

technical skills (at scientific and computational levels) and communication. These competencies will ensure a team has professional expertise in FAIR data handling. Almost all the participants stated that knowing how to create metadata and use ontologies in particular were necessary skills.

“I think data stewardship is really key. In addition, it would be infrastructure people who know how to set up a knowledge graph and how to maintain a knowledge graph. How to establish the ontologies—ontology is another central point.” (P1)

Investing in training as a facilitator of organisational culture and the subsequent implementation of FAIRification was also viewed as important. FAIRification was described as an emerging process in their companies, that required raising awareness and educating individuals about why they should adopt this new practice.

“It is an investment in training individuals; it’s not so much in the development of software systems.” (P2)

Another aspect that participants found important for promoting FAIRification was the existence of incentives, in particular for legacy data. Participants highlighted that the prevailing culture at the organisational level did not encourage retrospective FAIRification processes, as there was no incentive to do this and rather there was counter-pressure to meet the required productivity rates. They stated that the only incentive provided to them is encouragement from project managers or project teams.

“The legacy is always going to be an issue. How you’re going to push people to go back to their data and really make it FAIR, that’s going to be an issue, unless there’s some reward at the end.” (P11)

3.1.3 Theme 3: FAIRification Decision-making Process

This theme addresses how decisions are made about whether to FAIRify existing data sets. It covers prioritisation and resource allocation, ethical and legal considerations, and the role of management in the process.

(1) Data set prioritisation

Participants noted that prioritising legacy data for FAIRification is a complex process within R&D departments. They described the success of this task as primarily dependent on optimal resource use, which would in turn depend on capacity issues and the volume of legacy data. They emphasised that prioritisation is based on the data set’s value and relevance to each corresponding project.

“We have to be selective. The reason against would be we have a lot of legacy data, and people have to say that they’re interested in it, or someone has to make a decision that these data are valuable enough to invest in the work required for FAIRification.” (P6)

The participants also emphasised that a data set's uniqueness and competitive advantage make it a high priority for FAIRification. If the data set can be demonstrated to confer a competitive advantage for their organisation, this would make it a higher priority for the curation and re-annotation necessary to align it with FAIR principles.

“Whether the data set is actually proprietary to the organization. If it is a competitive advantage that will raise its profile and its priority for curation.” (P5)

The participants also identified the characteristics of a data set as a factor that plays a significant role in prioritisation. They tended to prioritise data sets according to their data quality: how complete the metadata were and whether they met existing standards.

“When I look at the characteristics of legacy data and whether it’s worth FAIRifying them or not, I tend to think along the lines of how complete is the existing metadata? Does it conform to an existing standard? What is the potential scientific or business impact?” (P5)

Participants also highlighted the importance of balancing the costs and benefits of data set FAIRification when making a prioritisation decision. This entailed estimating the resources required to FAIRify the legacy data and the expected need for the data.

“Within pharma, that’s similar in terms of the cost benefit... There is willingness, it’s just how much it’s going to cost and which data sets are worth it.” (P11)

The views of research scientists were important in identifying requirements and assessing the expected benefits.

“I think we rely on the research scientist to tell us what they need and what they would like. We can give them examples of the benefit of FAIRification, they say if they like it or not. They give us feedback.” (P6)

It was not always the case that FAIRification was viewed as the most cost-effective option. The cost of experiments is decreasing so dramatically that it may actually be more efficient to rerun an experiment using new types of instruments than to reuse existing data, unless the data are unique. Advances in technology mean the effort that goes into working with historical data is not necessarily worthwhile.

“You’re dealing with the advance of technology and the advance of the quality of data and the advance in the range, sensitivity, depths of analysis that you can perform. All of those things are driving against investing large amounts of effort into working with historical data”. (P2)

Some participants reported a drive towards generating new data sets in preference to maintaining historical data, particularly in genome sequencing, as the cost of re-sequencing is actually lower than the cost of managing existing data. They reasoned that new data sets would also allow more relevant data to

be obtained, along with better analysis due to advances in equipment and even gene editing to introduce or remove genes that might be relevant to a disease.

“...Maintaining the raw data files from genome sequencing is not cost-effective, Because the cost of re-sequencing is lower than the cost of maintaining the data archive.” (P7)

(2) Legal and ethical considerations

Legal and ethical issues are a major consideration in the decision about whether to FAIRify data. The legal aspect of access rights is a significant issue due to its complexity and the lack of clear process for accessing previously generated data sets, which may incur a significant cost to clarify. Legal complications are a particular challenge when multiple countries are involved, and each country has its own legislation with regard to access to legacy data.

“Sometimes ... you have all the legal aspects to get it or not, actually pretty expensive to clarify, if you can actually access these data. Some of these data are constrained with respect to what kind of consent you have. What can you do with this data set? The legal aspect is very complex especially for this FAIR, it doesn't really fit with the big and clear process.” (P10)

Another issue affecting accessibility is that a lot of data are generated through contract research organisations (e.g., service providers), who retain control over access and privacy issues in their research agreements.

“The access component we are still working on because it is very complicated in our area because of the research data agreements.” (P14)

Although it would be more efficient to FAIRify clinical data than conduct new studies, as this is a particularly expensive part of drug development, ethics compliance is an issue in the reuse of this type of data. The collection of clinical data involves considerable compliance processes, which may be subject to retrospective challenges with respect to regulations, audits and patient privacy. For example, if patients in the original study did not explicitly consent to sharing the data, then it may not be legally possible to reuse that data.

“Retrospective use of data from clinical trials can often be a problem, simply because the older informed consent from past clinical trials may not be drawn up in such a way that reuse of the information is actually possible.” (P7)

(3) Management commitment

The participants stated that decision making about retrospective FAIRification is a joint process between upper management and a data strategy team. They explained that the process requires interaction between these divisions as a managerial decision is required to approve data FAIRification, while a strategy team decides on how to make progress with the process.

“That would be a joint decision between the business leader of that domain and IT leader for the cost to do that FAIRification and the value that would return scientifically or corporately as a business/commercial function.” (P5)

Management is obliged to approve a particular FAIRification process, but this approval is based on a discussion with a team that is empowered to determine how to actually execute the process and with which data to commence.

“Then there is a team. Then that team proposes a route, how to get to a FAIR omics data landscape. This route then, obviously, is discussed with management and has been approved. This is how the process goes. There is a strategic decision to go into the field, and then the team decides how to deal with that field.” (P1)

The importance of having a long-term data strategy was raised as a critical factor for the enforcement of FAIR principles, with buy-in from the highest level necessary to executing this successfully.

“It is a bottom-up request but a top-down instruction, endorsement of following these processes. That doesn’t mean that all the departments, all the therapeutic areas, are very strongly aligned. There, I feel that within the company now, there is also a momentum shift looking to almost a president level decision on making sure all the different activity lines and strategies are following the same harmonised FAIRification approach.” (P13)

3.2 Conceptual Model

Although it is now more common to consider FAIR implementation in pharmaceutical R&D at the beginning of the project (FAIR by design), the FAIRification of legacy data remains a major focus. How decisions are made about retrospective FAIRification thus emerged as the primary concern in pharmaceutical R&D. This process has several steps, depends on many factors and involves various stakeholders, as illustrated in Figure 2. The cooperation between the management team (which may include the IT leader, middle manager, and lab head) and the data team (including data providers and producers (e.g., researchers), data consumers (e.g., data scientists), and data stewards) facilitates the process of selecting legacy data for FAIRification. The data team may begin by prioritising use cases (an example or experiment) and identifying related studies based on their relevance to each corresponding project. In other cases, the management team selects specific data for FAIRification.

This prioritisation of use cases is influenced first by factors such as ability to access the data set and the ethical governance requirements. Then, the data team assesses the effort required to FAIRify the data based on its characteristics (e.g., whether it meets existing standards, how old it is). At the same time, they also identify the benefits of FAIRification based on the value of the data set. The data team provides feedback to management who will ultimately approve FAIRification based on their assessment of both the scientific and the business case. If management approves the proposed FAIRification, the data team then defines the process and determines the scope of the project, and the requirements involved in FAIRifying the chosen

data set. Finally, management allocates resources (employees to do the work, a certain amount of time, etc.) to carry out the FAIRification. Once the FAIRification has been completed, the FAIRified data are approved again by the management, so the data team can add the data to the company data catalogue. There is also some time allocated for ad hoc processes in case any stakeholders (e.g., researchers, scientists) have a specific use case or a question that they need to be addressed.

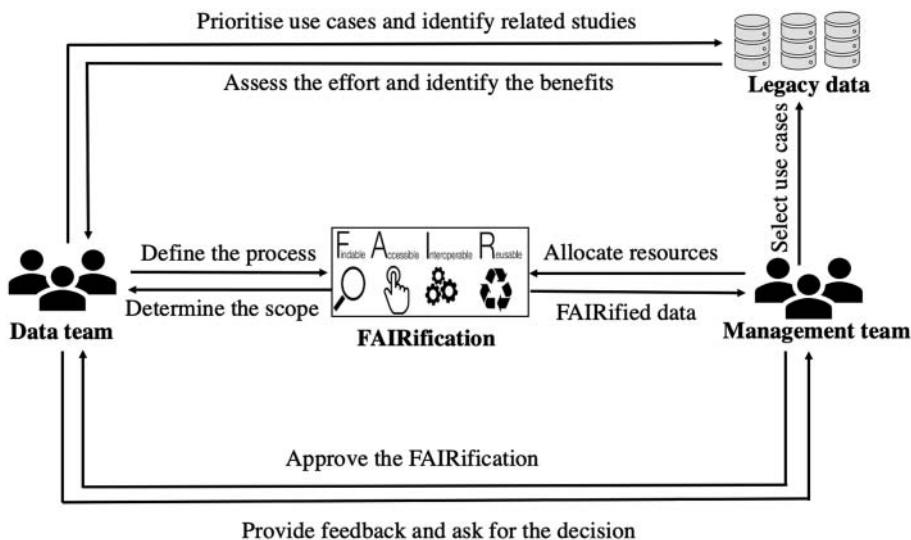


Figure 2. Conceptual model for the FAIRification decision-making process.

4. DISCUSSION

This study examined the implementation of FAIR principles in pharmaceutical R&D, through semi-structured interviews with 14 pharmaceutical professionals. The thematic analysis of the transcripts enabled us to gather insights about the practical realities of implementing the FAIR data principles in the field. Three primary themes emerged: FAIRification benefits, FAIRification costs and the FAIRification decision-making process. We found that adherence to FAIR guidelines can potentially improve drug R&D by generating current and future value from the reuse of data assets. Nevertheless, FAIRifying data entails considerable expenditure in terms of resources, both technical and human, along with training to promote cultural change. The decision-making process for retrospective FAIRification is complex, involving multiple teams and stakeholders, and requiring interaction between data scientists and management.

The findings reported here are supported by those of previous studies investigating the implementation of FAIR principles in the pharmaceutical industry [12, 13], which highlighted the expected benefits of the implementation of FAIR principles and the anticipated requirements for financial investment, cultural change, training and the technical infrastructure. Research has also highlighted the challenge of dealing with legacy data [30, 31]. FAIRifying data retrospectively remains challenging when data and metadata are

curated and re-annotated retrospectively [29]. We extend the literature by documenting another critical aspect of FAIRification—the decision-making process.

The reusability of data assets for the generation of further value was identified as the primary driver of FAIRification. This could enable repurposing of a drug that a company already has or uncover further uses or potential side effects which may not otherwise appear without further experimentation. For example, the availability of previous SARS-CoV-2 data presented in a FAIR format has contributed to an efficient response to the COVID-19 pandemic. In a similar vein, a recent study demonstrated that the availability of FAIRified primary genomic data could have helped the response to the pandemic [35]. To ensure an effective response to future outbreaks, several active communities have started defining the roadmap for FAIR implementation in health data [36, 37]. The respondents said that readily available data would enable large-scale analysis and powerful new AI analytics. These arguments are consistent with the findings of several studies that reported improved analysis owing to the availability of substantial high-quality, better-curated data [12]. In addition to effective analysis, advanced drug discovery processes are also enabled by the availability of well-managed data [38, 39, 40].

Culture change was noted as essential for the effective implementation of FAIR principles, in terms of raising awareness within an organisation of the potential value of FAIR data. Investment in training would be required to help people understand the value of reuse, as well as why data are an asset to companies. This cultural shift is expected to change people's perspectives of what they are valued for—that they are highly regarded not only for completing immediate project objectives, but also for creating valuable data sets for use in the future. An important consideration, however, is that there are prerequisites to adopting a FAIR culture, in particular, demonstrating how working in a FAIR-oriented manner generates long-term advantages and benefits, and being able to provide examples of FAIRification. Other studies have highlighted the importance of investing in culture change for the purpose of advancing the use of FAIR data in pharmaceutical companies [12], and that culture change is the principal obstacle to FAIR data implementation [41]. The respondents in this study stated that the only incentive provided to them at present is encouragement from project managers or project teams. The literature appears to have devoted little attention to incentivisation, and studies that do explore this matter have been conducted in academic, rather than industrial, contexts [42, 43, 44, 45].

Legal and ethical considerations are also important in the FAIRification decision-making process. Accessing legacy data can be complex, as the access request process often has an ad hoc design. This may be due to the fact that many of their data are generated by research organisations, which often retain control over access. While the value of reusing clinical data is clear, there may be ethical issues in terms of patient privacy and consent. Other studies have reported similar ethical challenges when it comes to implementing FAIR principles in human [46] and clinical [47] data. A recent review of FAIR data in health and medical research introduced additional principles to support compliance with legal requirements [14]. The authors showed how to resolve privacy and access challenges in handling health data, such as using privacy-enhancing technologies for anonymisation and minimising the risk of privacy breaches. Another study proposed using a FAIR-aware patient consent framework for data providers of human genomic data sets [48].

The current study has a number of limitations. The sample size was small due to a lack of participants who meet the inclusion criteria as FAIR implementation is a new area in pharmaceutical R&D. Participants were aware that they were being interviewed by a researcher focused on the development of a decision-making framework for FAIRification to be used by stakeholders in pharmaceutical R&D, which may have biased their responses. Moreover, when selecting participants, we targeted pharmaceutical professionals with different levels of experience and involvement in various aspects of FAIRification. This means that our sample comprised senior-level participants who were involved in management decision-making and other employees with prior FAIRification experience. Respondents' perspectives might have differed if they had been implementing FAIRification as part of their day-to-day business. As with all interview studies, the results may not generalize to other samples or populations.

5. CONCLUSION

This study examined the implementation of FAIR in pharmaceutical R&D, which is a new practice in many pharmaceutical organisations. We found that the implementation of these guiding principles as a form of cooperative data management has the potential to increase the higher reusability of data assets and significantly reduce costs of drug discovery and development. Nevertheless, it remains the case that retrospective FAIRification in particular entails significant costs, and that a culture shift is required to support its implementation. One of the significant findings to emerge from this study is the identification of the process of how the decision about retrospective FAIRification is made.

ACKNOWLEDGEMENTS

C. Goble and N. Juty acknowledge the FAIRplus project (IMI2 Joint Undertaking under grant agreement No. 802750). The EPSRC supported C. Jay in this work under EP/S021779/1. E. Alharbi's scholarship is sponsored by Umm Al-Qura University, the Kingdom of Saudi Arabia (No. 1057493924).

AUTHOR CONTRIBUTIONS

E. Alharbi (ebtisam.alharbi-3@postgrad.manchester.ac.uk), C. Goble (carole.goble@manchester.ac.uk), C. Jay (caroline.jay@manchester.ac.uk), and N. Juty (nick.juty@manchester.ac.uk) initiated the effort and conceived the paper. E. Alharbi wrote the first draft of the manuscript, collected and analysed the data. R. Skeva (rigina.skeva@manchester.ac.uk) analysed the data. C. Goble, C. Jay, and N. Juty provided critical ideas for the research undertaken and also provided supervision and feedback to help shape the manuscript. All authors contributed to reviewing and editing of the final version of the article.

REFERENCES

[1] Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, Article number 160018 (2016)

[2] Mons, B., et al.: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. *Information Services & Use* 37(1), 49–56 (2017)

[3] European Commission: Turning FAIR into reality. Available at: <https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283>. Accessed 21 September 2021

[4] Bloemers, M., Montesanti, A.: The FAIR funding model: Providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices. *Data Intelligence* 2(1–2), 171–180 (2020)

[5] Velterop, J., Schultes, E.: An academic publishers' GO FAIR implementation network (APIN). *Information Services & Use* 40(4), 333–341 (2020)

[6] Jacobsen, A., et al.: FAIR principles: Interpretations and implementation considerations. *Data Intelligence* 2(1–2), 10–29 (2020)

[7] Mons, B.: Data stewardship for open science: Implementing FAIR principles. 1st edition. Chapman and Hall/CRC, New York (2020)

[8] European Commission: Cost-benefit analysis for FAIR research data—Cost of not having FAIR research data (2019). Available at: <https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en>. Accessed 21 September 2021

[9] European Commission: Realising the European Open Science Cloud (2016). Available at: <https://op.europa.eu/en/publication-detail/-/publication/2ec2eced-9ac5-11e6-868c-01aa75ed71a1/language-en>. Accessed 21 September 2021

[10] G7. Expert Group on Open Science (2017). Available at: <http://www.g8.utoronto.ca/science/2017-annex4-open-science.html>. Accessed 21 September 2021

[11] GO FAIR. Available at: <https://www.go-fair.org>. Accessed 21 September 2021

[12] Wise, J., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discovery Today* 24(4), 933–938 (2019)

[13] van Vlijmen, H., et al.: The need of industry to go FAIR. *Data Intelligence* 2(1–2), 276–284 (2020)

[14] Holub, P., et al.: Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health. *Biopreservation and Biobanking* 16(2), 97–105 (2018)

[15] Mestre-Ferrandiz, J., Sussex, J., Towse, A.: The R&D cost of a new medicine (2012). Available at: https://www.researchgate.net/profile/Jorge-Mestre-Ferrandiz/publication/290441583_The_RD_cost_of_a_new_medicine/links/5698e65808aec79ee32cacb9/The-R-D-cost-of-a-new-medicine.pdf. Accessed 21 September 2021

[16] Scannell, J.W., et al.: Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* 11(3), 191–200 (2012)

[17] Wise, J., et al.: The positive impacts of real-world data on the challenges facing the evolution of biopharma. *Drug Discovery Today* 23(4), 788–801 (2018)

[18] Vaudano, E.: The innovative medicines initiative: A public private partnership model to foster drug discovery. *Computational and Structural Biotechnology Journal* 6(7), e201303017 (2013)

[19] Blackburn, M., et al.: Big data and the future of R&D management: The rise of big data and big data analytics will have significant implications for R&D and innovation management in the next decade. *Research-Technology Management* 60(5), 43–51 (2017)

[20] Tormay, P.: Big data in pharmaceutical R&D: Creating a sustainable R&D engine. *Pharmaceutical Medicine* 29(2), 87–92 (2015)

[21] The Pistoia Alliance FAIR Toolkit. Available at: <https://fairtoolkit.pistoiaalliance.org>. Accessed 21 September 2021

[22] The IMI FAIRplus FAIR Cookbook. Available at: <https://github.com/FAIRplus/the-fair-cookbook>. Accessed 21 September 2021

[23] Kruhse-Lehtonen, U., Hofmann, D.: How to define and execute your data and AI strategy. *Harvard Data Science Review* 2.3 (2020). Available at: <https://hdsr.mitpress.mit.edu/pub/4vlrf0x2/release/1>. Accessed 21 September 2021

[24] Makarov, V., et al.: Best practices for artificial intelligence in life sciences research. *Drug Discovery Today* 26(5), 1107–1110 (2021)

[25] Fleming, N.: How artificial intelligence is changing drug discovery. *Nature* 557(7706), S55–S57 (2018)

[26] Vamathevan, J., et al.: Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18(6), 463–477 (2019)

[27] Slate, T.: Overcoming the challenges to making data FAIR in pharma (2020). Available at: <https://pharmafield.co.uk/opinion/overcoming-the-challenges-to-making-data-fair-in-pharma/>. Accessed 21 September 2021

[28] Jacobsen, A., et al.: A generic workflow for the data FAIRification process. *Data Intelligence* 2(1–2), 56–65 (2020)

[29] Rocca-Serra, P., Sansone, S.A.: Experiment design driven FAIRification of omics data matrices, an exemplar. *Scientific Data* 6, Article number 271 (2019)

[30] Genomics, F.: Driving FAIR in biopharma report (2021). Available at: <https://info.frontlinegenomics.com/driving-fair-in-biopharma>. Accessed 21 September 2021

[31] Front Line Genomics. Transforming R&D with data report (2020). Available at: <https://frontlinegenomics.com/transforming-rd-with-data-report/>. Accessed 21 September 2021

[32] Patton, M.Q.: Qualitative research. In: Everitt, B., Howell, D. (eds.) *The Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Chichester (2005)

[33] Silverman, D.: Qualitative research. SAGE, London (2020)

[34] DiCicco-Bloom, B., Crabtree, B.F.: The qualitative research interview. *Medical Education* 40(4), 314–321 (2006)

[35] Schriml, L.M., et al.: COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data* 7, Article number 188 (2020)

[36] Mons, B.: The VODAN IN: Support of a FAIR-based infrastructure for COVID-19. *European Journal of Human Genetics* 28(6), 724–727 (2020)

[37] Research Data Alliance (RDA) COVID 19 Working Group. RDA COVID19 case statement (2020). Available at: <https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid19-epidemiology-rda-covid19-clinical-rda-covid19-social>. Accessed 21 September 2021

[38] Chen, B., Butte, A.: Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology & Therapeutics* 99(3), 285–297 (2016)

[39] Lo, Y.C., et al.: Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* 23(8), 1538–1546 (2018)

[40] Brown, N., et al.: Big data in drug discovery. *Progress in Medicinal Chemistry* 57(1), 277–356 (2018)

[41] Stall, S., et al.: Make scientific data FAIR. *Nature* 570, 27–29 (2019)

[42] Tenopir, C., et al.: Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE* 10(8), e0134826 (2015)

- [43] Samota, E.K., Davey, R.P.: Knowledge and attitudes among life scientists towards reproducibility within journal articles. *BioRxiv* preprint BioRxiv: 581033 (2019)
- [44] Chawinga, W.D., Zinn, S.: Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research* 41(2), 109–122 (2019)
- [45] Borgman, C.L.: Data sharing and reuse in interdisciplinary scientific collaborations: Challenges of heterogeneous practice (2018). Available at: <https://escholarship.org/uc/item/6tb5718z>. Accessed 21 September 2021
- [46] Boeckhout, M., Zielhuis, G.A., Bredenoord, A.L.: The FAIR guiding principles for data stewardship: Fair enough? *European Journal of Human Genetics* 26(7), 931 (2018)
- [47] Lo, B.: Sharing clinical trial data: Maximizing benefits, minimizing risk. *JAMA* 313(8), 793–794 (2015)
- [48] Corpas, M., et al.: A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Computational Biology* 14(3), e1005873 (2018)

AUTHOR BIOGRAPHY



Ebtisam Alharbi is a Postgraduate Researcher with experience in managing information for the business. Aspires to continue research into FAIR data management in the pharmaceutical industry, with specific interests in pharmaceutical R&D. She is qualified in Computer Science (BA) and Information Management (MSc).

ORCID: 0000-0002-3887-3857



Rigina Skeva is a Postgraduate Researcher with experience of designing and evaluating Virtual Reality Therapy (VRT) for Substance Use Disorders. Aspires to continue research into VRT for treating psychological conditions, with specific interests in the therapeutic potential of avatar use (embodiment, social interaction, change of perspective). She is qualified in Primary Education (BA) and in Computer Science and Information Technology (MSc).

ORCID: 0000-0003-3816-4847



Nick Juty is a Senior Research Technical Manager. He is an experienced senior scientist with recent focus on standards adoption across scientific domains. He has played a leading role in delivering an international and cross-disciplinary identification system for scientific data (<http://identifiers.org>).

ORCID: 0000-0002-2036-8350



Caroline Jay is a Professor of Computer Science and joint Head of Research in the School of Engineering, University of Manchester. She is qualified as both a Psychologist (BA, CPsychol) and Computer Scientist (MSc, PhD), and undertakes research crossing these domains. She is Research Director of the Software Sustainability Institute (<https://www.software.ac.uk/>), and a keen advocate for open and reproducible science.

ORCID: 0000-0002-6080-1382



Carole Goble is a Professor of Computer Science at University of Manchester and deputy director EU ESFRI ELIXIR UK Node. In 2001 she established and co-directed myGrid, a sub-group of Information Management Group, which focuses on data intensive e-Science. The group ranges from theory to practice, translating state of the art techniques in semantic Web, distributed computing, data management and social computing into software and resources widely used by scientists from many different communities. The team is made up of scientific informaticians, computer science researchers and software engineers. We collaborate with scientists world-wide, from many disciplines: Life Sciences, Biodiversity, Astronomy, Chemistry, Health informatics, Social Science and Digital Libraries. In 2010 she co-founded the Software Sustainability Institute (<https://www.software.ac.uk/>).

ORCID: 0000-0003-1219-2137